CABIOS

Vol.2, no.2, 1986
Pages 99 – 101

SEQGEL: a versatile and comfortable DNA editor which supports a special keyboard and a speech synthesizer

Thomas R.Bürglin

Abstract

A DNA editor for an Apple II is described which contains many additional functions apart from just editing sequences. The data files are normal ASCII text or binary files and can thus be used easily by other programs. The program supports a special keyboard which greatly facilitates typing of DNA sequences. Furthermore a speech synthesizer is supported by the editor. The speech feedback, together with the special keyboard, reduces typing errors to a minimum.

Introduction

It is obvious that only larger computer systems can handle the increasing number of DNA and protein sequences, allowing searches of data banks with reasonable speed. However, simple tasks like entering a sequence do not require fast CPUs and rather block the terminals for other uses. Low-cost personal computers can perform such tasks very well.

In our institute DNA sequences were usually compiled on paper before they were typed into the computer. Thus much time was spent in writing and proofreading the sequences. Obviously a more attractive and convenient editing system for sequences was needed.

None of a collection of programs (Söll and Roberts, 1984) offered such a program for an Apple II system, so a flexible DNA editor program was written. The data format is compatible with the programs by Staden (1977). The data files can also be used directly or after one conversion with other published programs for the Apple II (Larson and Messing, 1982; Paollela, 1985; Dardel, 1985).

Two low-cost hardware devices which are supported by the program improve the speed and reduce the error rate when entering sequences. One is a special keyboard with only T,C,G,A and some additional keys. The second feature is a speech synthesizer.

System and methods

Hardware

The program was written on an Apple IIe and uses the following peripherals: 80 column card; two disk drives; one EPSON FX-80 printer with EPSON parallel interface (slot 1); one BROTHER daisy-wheel printer with EPSON parallel interface (slot 2); SUPER-SPEECH speech synthesizer (slot 7) (obtained from ECKL Electronic, Erlenmeyerstrasse 3, D-6204 Taunusstein 4, FRG for ~$60). For data transfer a Super Serial Interface is installed and linked to a VAX. However, not all the peripherals need be present. The minimal requirement is an Apple II+ or Apple IIe with two disk drives. Some functions in the PRINT section of the program have to be modified, depending on the connected peripherals.

For convenient typing a special keyboard was designed which contains only the necessary characters for entering a DNA sequences and some additional keys for easy editing (Figure 1). The keyboard is connected in parallel with the Apple keyboard. Figure 2 shows which lines have to be used for connecting the keyboard.

Language and data files

The main program is written in APPLESOFT BASIC. Time-consuming operations are written in 6502 machine code. The program runs under DOS 3.3. The data files created by the program are either text files with a carriage return after every 60 characters or binary files without any intervening characters in the sequence. When reading in text files, the program will remove any control characters and convert most of the special characters to a '–'.

A version running under PRODOS is also available. The data files created are only text files without any intervening characters. However, other text files containing control characters can be read. The sequences in all the data files are normal ASCII characters. No header or other control information is present.

The program can also create text files (so called CATALOG files, similar to the FILE of FILES described by Staden, 1977), which contains the filenames of sequences. Thus it is possible to keep lists of related sequences.

Sequences up to 12400 characters can be handled by the program.

Program description

The program uses the 40 column screen for better visibility of the characters. The screen is split into two parts. The top part has four lines as a header: the current position (C.P.) and length

Department of Cell Biology, Biocenter of the University of Basel, CH-4056 Basel, Switzerland

Present address: Department of Biological Chemistry, UCLA School of Medicine, Los Angeles, CA 90024, USA

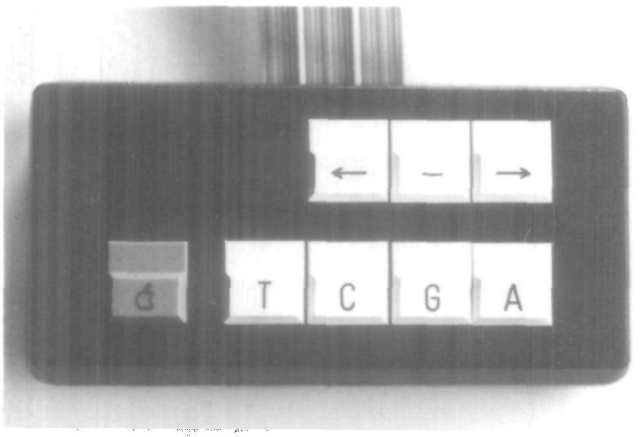

Fig. 1. Picture of the special keyboard with T,C,G,A the arrows for moving back and forth in the sequence, the dash for unknown nucleotides and the 'OPEN-APPLE' key, which gives double functions to the other keys. It is ~ 13 cm wide.

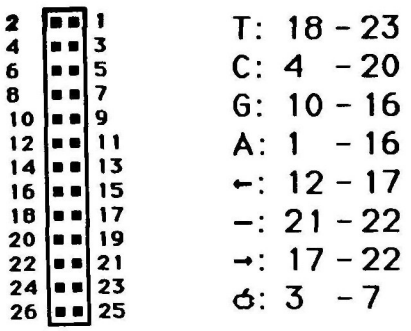

Fig. 2. Scheme of the keyboard plug with numbering for the Apple IIe. The numbers shown have to be connected to the corresponding key.

are displayed at the top. Then there is a window of the sequence from −15 to +15 of the c.p. The next line is for entering commands. The bottom part is a normal scrolling screen.

The following commands are available:

E (ENTER/OVERWRITE MODE) for entering and editing a sequence.
I (INSERT MODE) for insert characters.
V (VERIFY MODE) for verifying a sequence.
D (DELETE) for deleting within a specified range.
DS (DELETE START) marks the start position of a sequence range to be deleted.
DE (DELETE END) marks the end position of a sequence range to be deleted.
L (LIST) lists the sequence on the screen.
P (PRINT) prints the sequence.
TRANS (TRANSLATE) translates a DNA sequence into single-letter amino acid code on disk.
G (GO TO) jumps to the specified position in the sequence.
T (TOP) jumps to the beginning of the sequence.
B (BOTTOM) jumps to the end of the sequence.
F (FIND) searches for a specified string.
COM (COMPLEMENT) complements the sequence.

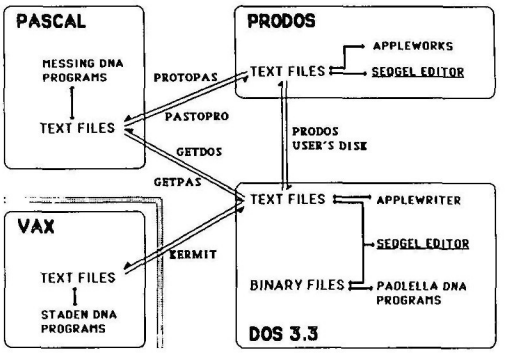

Fig. 3. The scheme illustrates how the different programs and operating systems relate to each other.

REV (REVERSE) inverts the sequence.
RC (REVERSE COMPLEMENT) reverse complements the sequence.
NOERR (NO ERROR) replaces all small letter nucleotides (used to indicate uncertainties) with their capital counterparts.
SP (SPEECH) switches the speech on and off. When speech is on, the following sounds will be generated: T, C, G, A, X for '–', uncertain nucleotides will be preceded by a MAYBE. Mismatches and errors will cause a MISTAKE.
NEW (NEW) for entering a new sequence.
EX (EXIT).

Commands for the floppy disk:

CATALOG makes a list of all files on the disk.
SAVE saves the file (sequence) on disk.
LOAD loads the file into the computer memory.
APPEND joins two sequences.
DELETE deletes a disk file.
LOCK protects a disk file from deletion.
UNLOCK deprotects a disk file.
CFLIST (CATALOG FILE LIST) list the Catalog file.
CLR (CLEAR) clears the filenames in memory.

In the ENTER, INPUT and VERIFY modes one can use the normal and the special keyboard for entering and checking sequences. With the arrows you move back and forth through the sequence. Pressing 'open-apple' together with T, C, G or A produces lower case letters which can be used for representing uncertain nucleotides. 'Open-apple' together with the arrows moves you automatically through the sequence. While you are in these command modes, you always get speech feedback.

The PRINT section allows you to print the sequences in various formats onto printers or disk (for word processing) either as single-stranded, double-stranded or translated version.

Discussion

The DNA editor described here is fast and convenient to use due to the command structure and the machine language routines. In addition, no reloading of sequences or program

parts is necessary when different functions are used. The format of the data files was chosen so that many other programs can use them. The binary files created by the program can be used directly by the programs described by Paolella (1985). The text files can be used with the programs of Dardel (1985). Alternatively they can be converted to PASCAL, where they can be used with programs described by Larson and Messing (1982) or with a modified version of the programs described by Malthiery *et al.* (1984). Furthermore, they can be transferred to a VAX computer where they can be used with the programs described by Staden (1977).

Apart from the fact that the text files can be used directly in word processors like APPLEWRITER or APPLEWORKS, the print option allows printing onto the disk. Thus sequences can be transferred to word processors in numbered and translated versions, which eliminates another source of error. Figure 3 shows how the files and operating systems are related.

The use of the character '−' for unknown nucleotides is according to Staden (1977). It is generally accepted because of its better visibility than 'N' in sequences (Cornish-Bowden, 1985). The uncertainty codes used by Staden (1977) are useful for computer programs, but hard to remember. The use of lower case letters for uncertain nucleotides as used in this program is not standard. However, the significance of the small letters is obvious and they are easily detected in printouts. They are a reminder that there are uncertainties in this region of the sequence which have to be checked.

The special keyboard described offers a convenient way of typing in sequencing gels directly. The convenience is certainly similar to digitizing tablets which have been described for other programs (Söll and Roberts, 1984), but the costs are much lower. In addition, the keyboard is also very convenient for typing in printed sequences.

The program supports a speech synthesizer. This is extremely valuable for initial entering as well as for checking of sequences. The computer speaks the sequence, while a single person can control it on a gel or listing. Because the main part of the program is written in BASIC, modifications for other printers or speech synthesizers should be easy.

Acknowledgements

I would like to thank Jürg Wenger and Paul Henz for building the keyboard, and Bernard Jacq, Erich Frei and Denice Smith for critical reading of the manuscript.

References

Cornish-Bowden,A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **13**, 3021–3030.

Dardel,F. (1985) PEGASE: a machine language program for DNA sequence analysis on Apple II microcomputer using a binary coding of nucleotides. *CABIOS*, **1**, 19–22.

Larson,R. and Messing,J. (1982) Apple II software for M13 shotgun DNA sequencing. *Nucleic Acids Res.*, **10**, 39–50.

Malthiery,B., Bellon,B., Giorgi,D. and Jacq,B. (1984) Apple II Pascal programs for molecular biologists. In Söll,D. and Roberts,R.J. (eds), *The Applications of Computers to Research on Nucleic Acids II, Part 2*. IRL Press, Oxford, UK, pp. 569–579.

Paolella,G. (1985) A fast DNA sequence handling program for Apple II computer in BASIC and 6502 assembler. *CABIOS*, **1**, 43–49.

Söll,D. and Roberts,R.J. (1984) *The Applications of Computers to Research on Nucleic Acids II, Part 2*. IRL Press, Oxford, UK, 428 pp.

Staden,R. (1977) Sequence data handling by computer. *Nucleic Acids Res.*, **4**, 4037–4051.

Received and accepted on 10 March 1986